REPORT

HANSHIN UNIVERSITY

과목명 | 빅데이터분석실습

담당교수 | 이승천 교수님

학과 | 응용통계학과

학년 | 4학년

학번 | 201452024

이름 | 박상희

제출일 | 2019. 04. 10

Q. Sonar 데이터를 Logistic Regression, neural network으로 분석하고 20%의 test 데이터에 대해 오분류표와 ROC Curve를 구하라.

==========

01. 데이터 설명

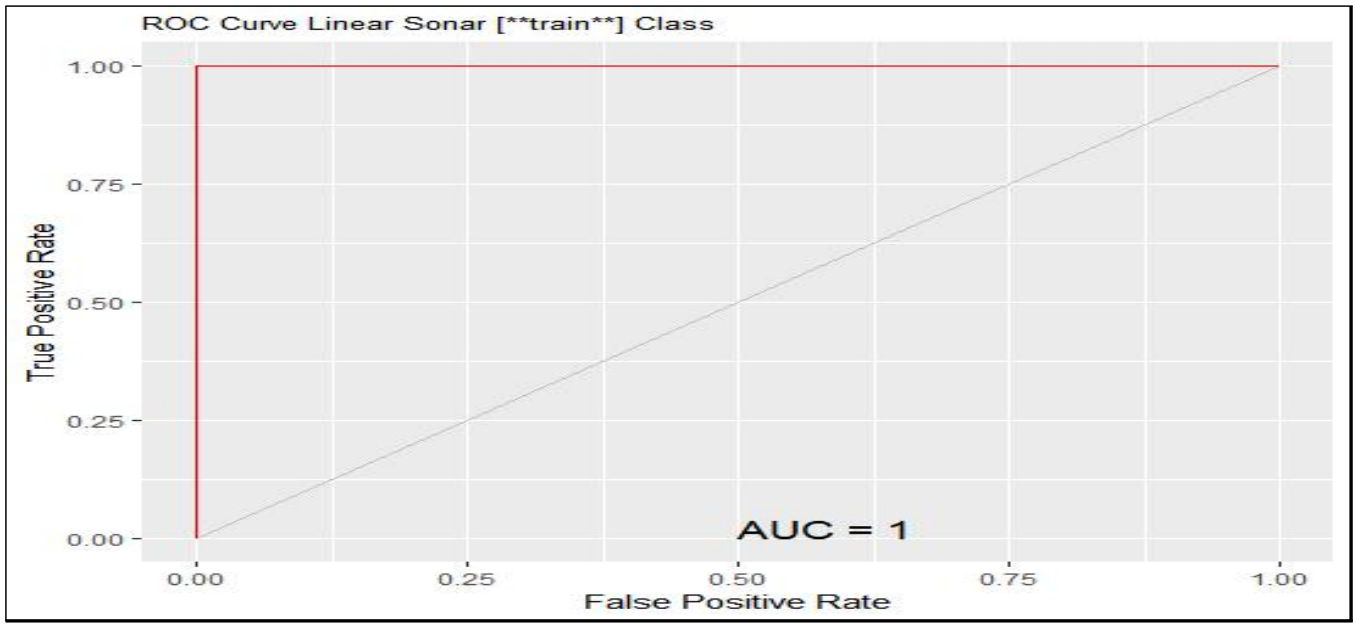
- 208 obs, 61 variables
- V1 ~ V60 : 0 ~ 1 사이의 numeric variables
- Class : M과 R로 이루어진 factor variable

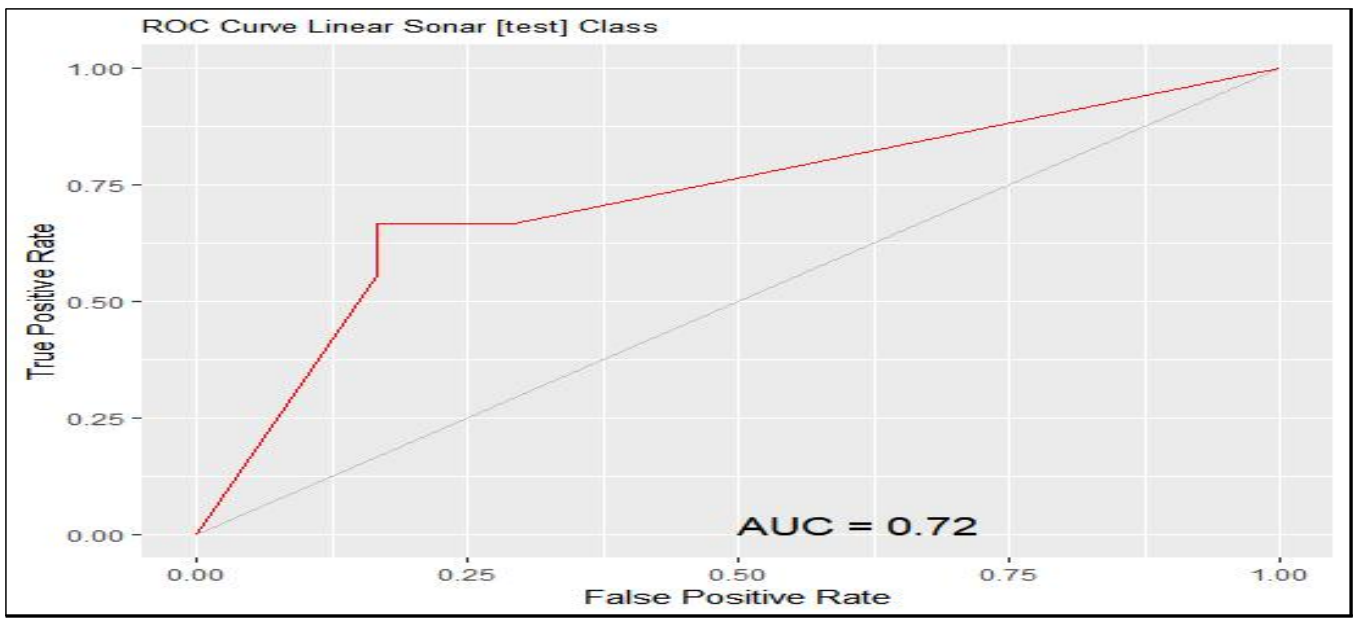
02. Logistic Regression

- Data = Train Data(80%, 166 obs) + Test Data(20%, 42 obs)

Train Data (166 obs)		Predicted	
		M	R
Actual	M	80	0
	R	0	79

Test Data (42 obs)		Predicted	
		M	R
Actual	M	20	4
	R	6	12



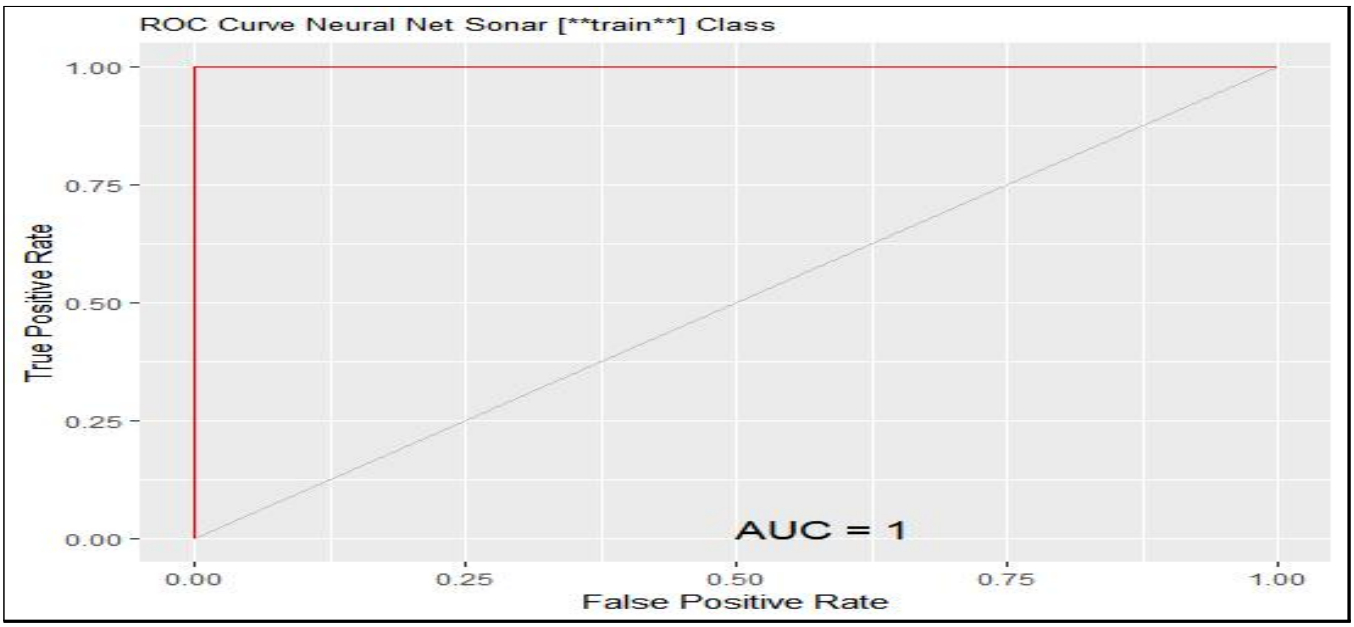


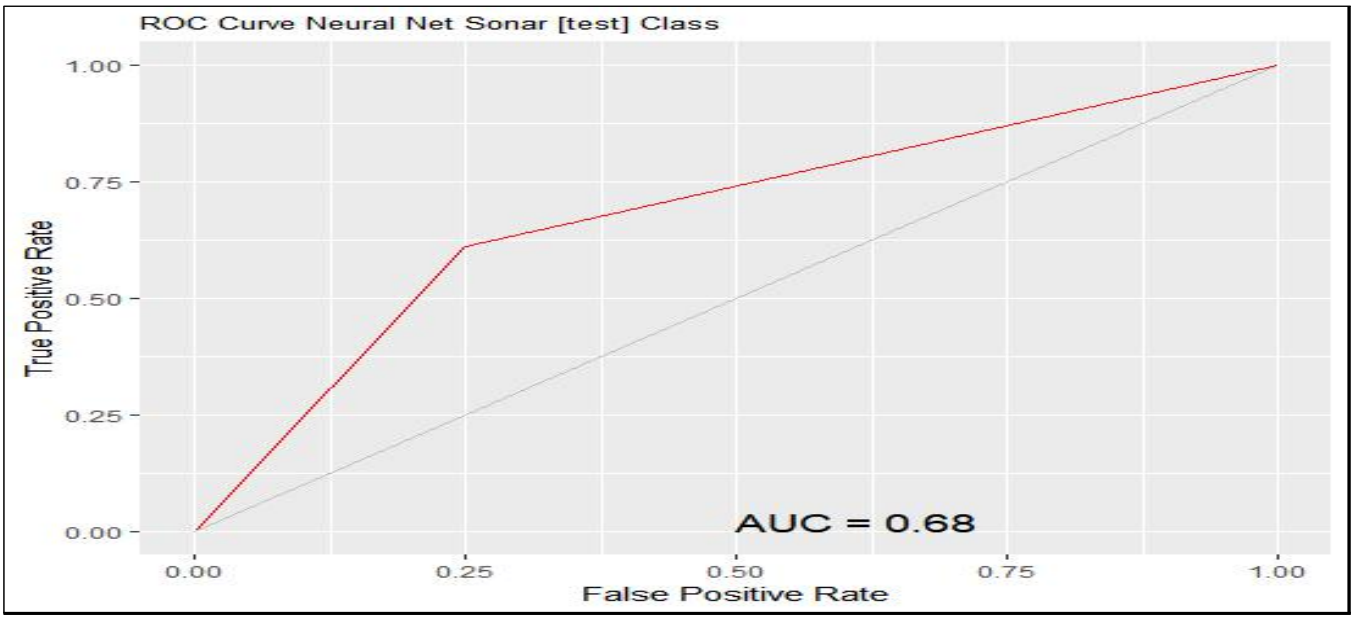
03. Neural Network

- Hidden Layer Nodes : 3
- Data = Train Data(80%, 166 obs) + Test Data(20%, 42 obs)

Train Data (166 obs)		Predicted	
		M	R
Actual	M	87	0
	R	0	79

Test Data (42 obs)		Predicted	
		M	R
Actual	M	18	6
	R	7	11





Q. Mroz 데이터에 대해 신경망 모형을 적합시키되, test 데이터를 기준으로 최적의 오분류율을 갖는 hidden node의 수를 탐색하라.

==========

01. 데이터 설명

- 753 obs, 8 variables
- lfp : 결혼한 백인 여성의 직업 여부 (factor)
- inc : 여성을 제외한 가구 수입 (numeric)
- wc : 여성의 대학 졸업 여부 (factor)
- hc : 남편의 대학 졸업 여부 (factor)
- lwg : 여성의 예상 임금 (integer)
- age : 나이 (integer)
- k5 : 5세 이하의 자녀의 수 (integer)
- k618 : 6~18세 이하의 자녀의 수 (integer)

02. 그리드 서치 (Grid Search)

- Data = Train Data(80%, 602 obs) + Test Data(20%, 151 obs)
- n : Hidden Layer Nodes
- AUC : Test Data(20%)에 대한 AUC 면적
- 정분류율 : Test Data(20%)에 대한 맞힌 개수 / 전체 개수
- n<5에서 AUC 값이 0.71이며, 정분류율이 0.70으로 가장 높은 n=2을 선택

n	1	2	3	4	5	6	7	8	9	10
AUC	0.72	0.72	0.71	0.65	0.69	0.69	0.5	0.83	0.81	0.69
정분류율	0.70	0.70	0.68	0.64	0.65	0.70	0.40	0.71	0.70	0.69